UNIVERSITY OF CAMBRIDGE

Laboratory for Scientific Computing

Random recursive tree ensembles: A high energy physics application

Outline

- The physics goal of this project is to classify Higgs to tau-tau decays (signal process) from background decays.
- The classification task aims to improve the statistical significance of the existence hypothesis [1].

  H_0: Only background processes exist.

  H_1: The signal process $H \rightarrow \tau^+\tau^-$ exists.
- The significance of the alternative hypothesis is quantified by the metric *Approximate Median Significance* (σ) which is another way of expressing a *p*-value.
- A significance of 5σ is required to claim a discovery of a new decay. The Higgs to tau-tau analysis has not yet achieved 5σ hence, the decay is unobserved in nature.
- Classifying signal from background is a challenging task since the classes completely overlap, the signal process is embedded within a dense background and is largely inseparable.

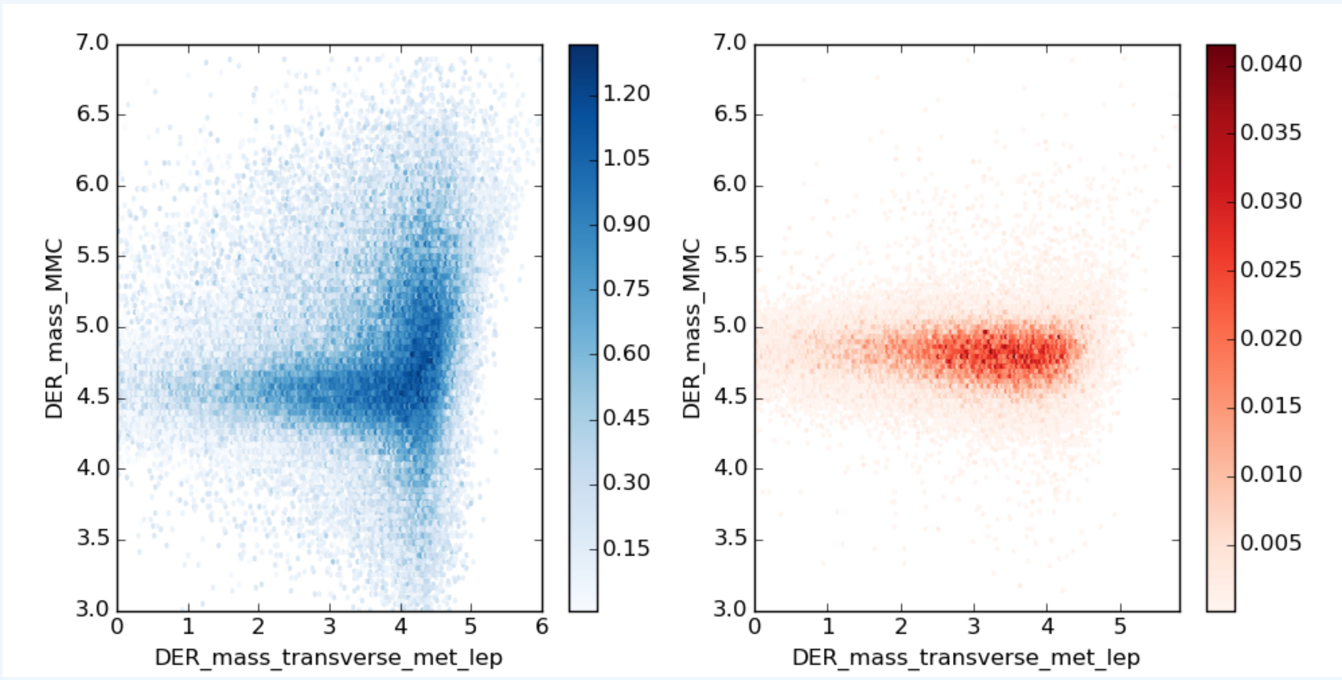
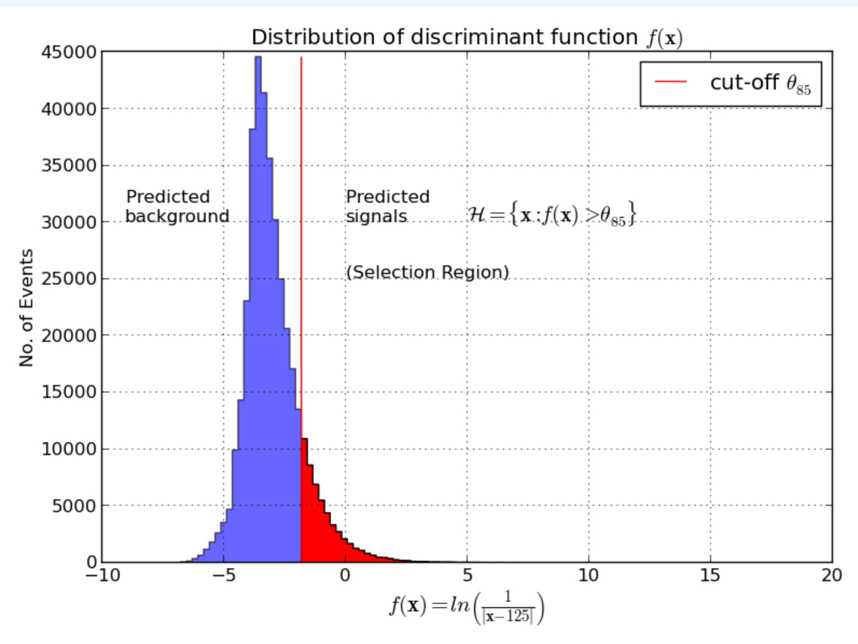


Figure 1: Background and Signal in 2d feature space

- In this project we propose a meta algorithm for the binary classification task and measure its performance through the AMS (σ) metric.



$$AMS = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

The AMS is computed on the basis of events in the selection region of a classifier.

Tree learning

The primitive binary tree is constructed by applying axis-parallel splits on each feature until a stopping criterion is reached. This gives rectangular decision boundaries. When trees are combined these rectangular regions intersect giving more intricate boundaries.

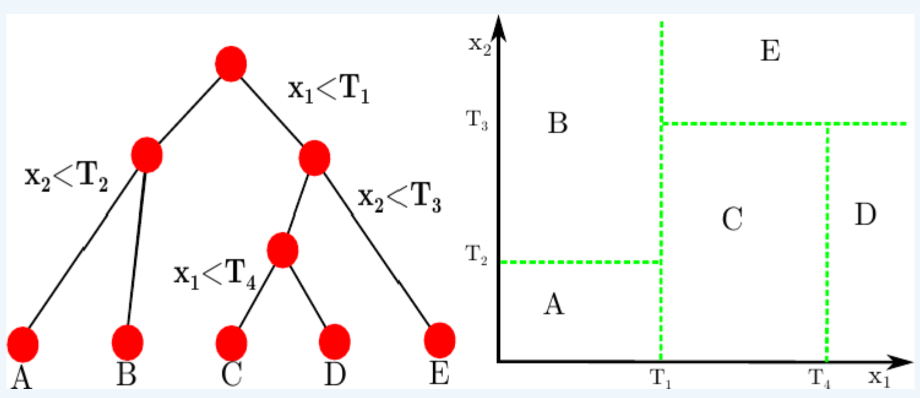


Figure 2: Depiction of the rudimentary tree learner in 2 features.

The tree (left) has a depth of 4 and the nodes [A,. . .,E] represent leaves, these might not be pure (i.e. contain samples of one class). All samples end up in a leaf and are assigned a score equal to the fraction of events of the same class in that leaf.

For example, if leaf C has 30 background events and 2 signal events, the signal events in leaf C will have a score of $1/16$ and background events will have a score of $15/16$.

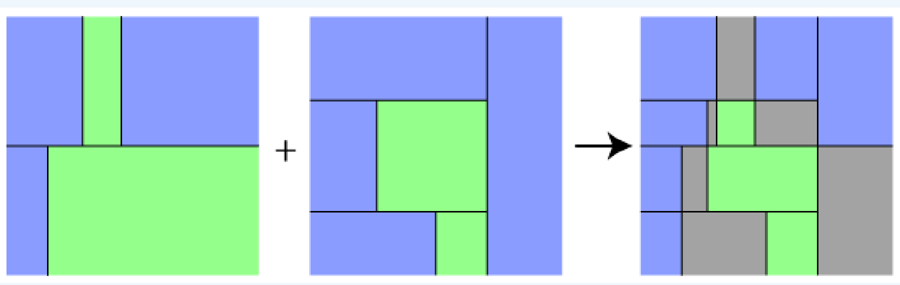


Figure 3: Tree boundary of a 2 tree forest

Algorithm

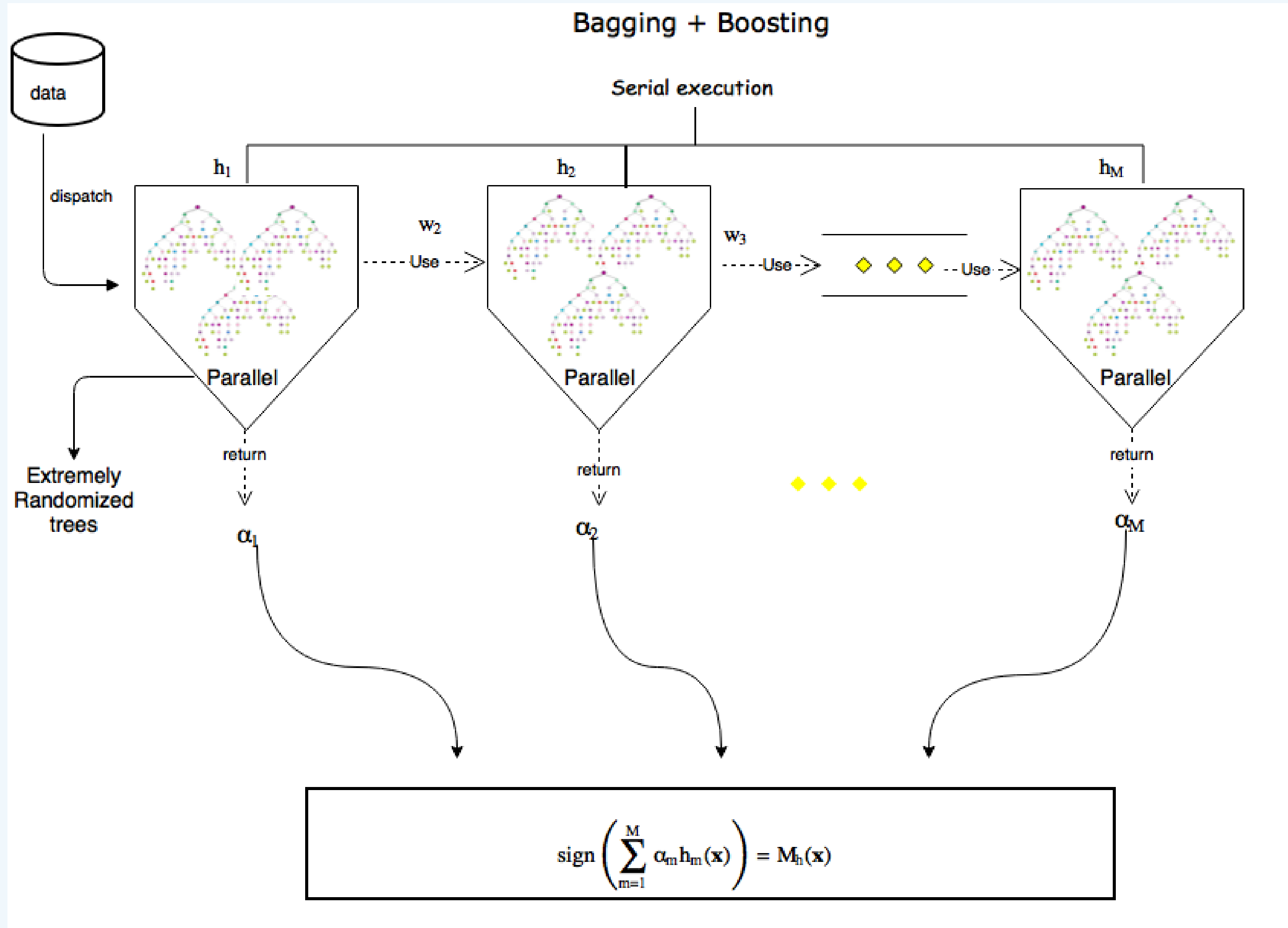


Figure 4: Meta-algorithm that combines bagging and boosting

Algorithm 1: Splitting algorithm used in extremely random trees

1. **Select** K features $x_1, \ldots, x_K$ at random from the training set $\mathbf{D}$.
2. **Select** K splits $\{T_1, \ldots, T_K\}$, one per feature for the K features chosen in step 1; each T_i is selected at random from the range of the feature values $\forall i = 1, \ldots, K$.
3. **Rank** the splits T_i by a criterion say Q which gives a score $Q(\mathbf{D}, T_i) \in \mathbb{R}$ for each split.
4. **Return** $T_* = max_{i=1 \ldots K} Q(\mathbf{D}, T_i)$.

Results

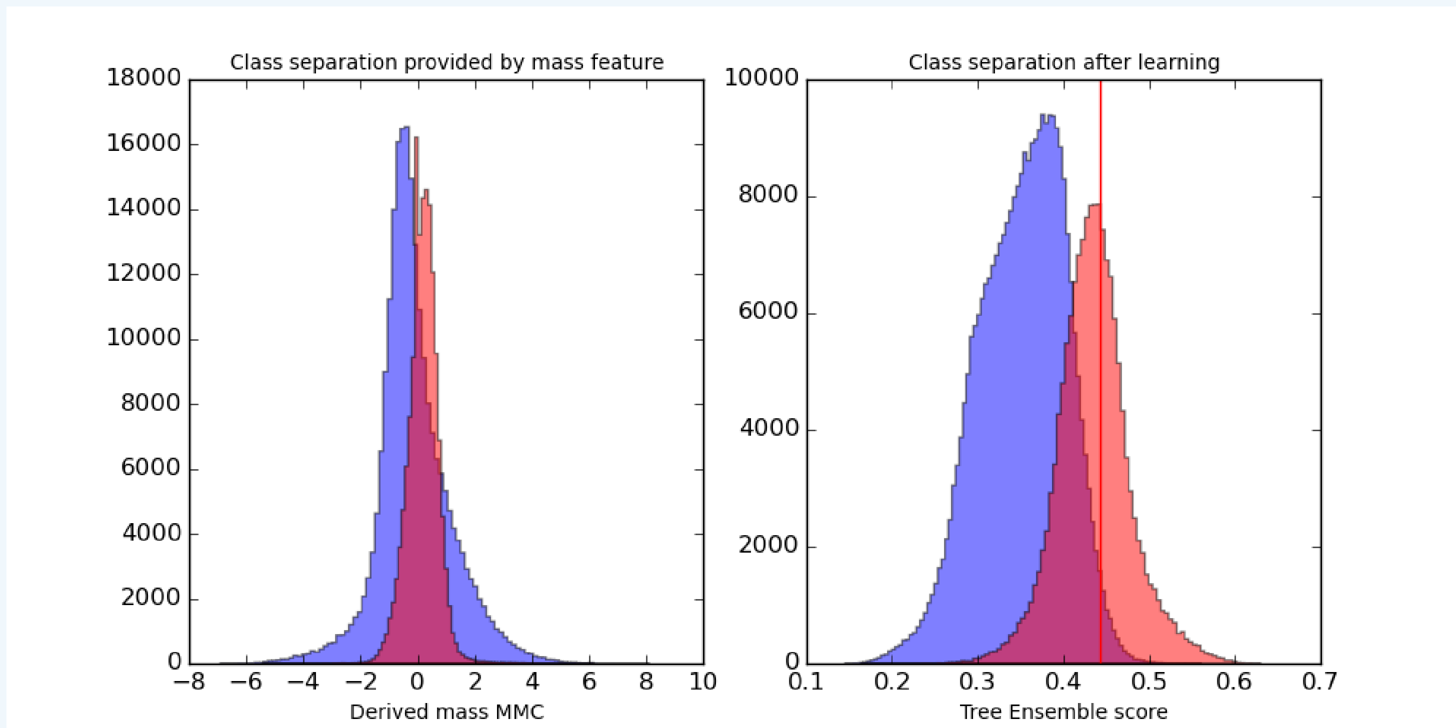


Figure 5: Signal and background separation provided by the mass feature (left) and by the tree ensemble score (right). The vertical line defines the selection region. We can notice superior separation post learning.

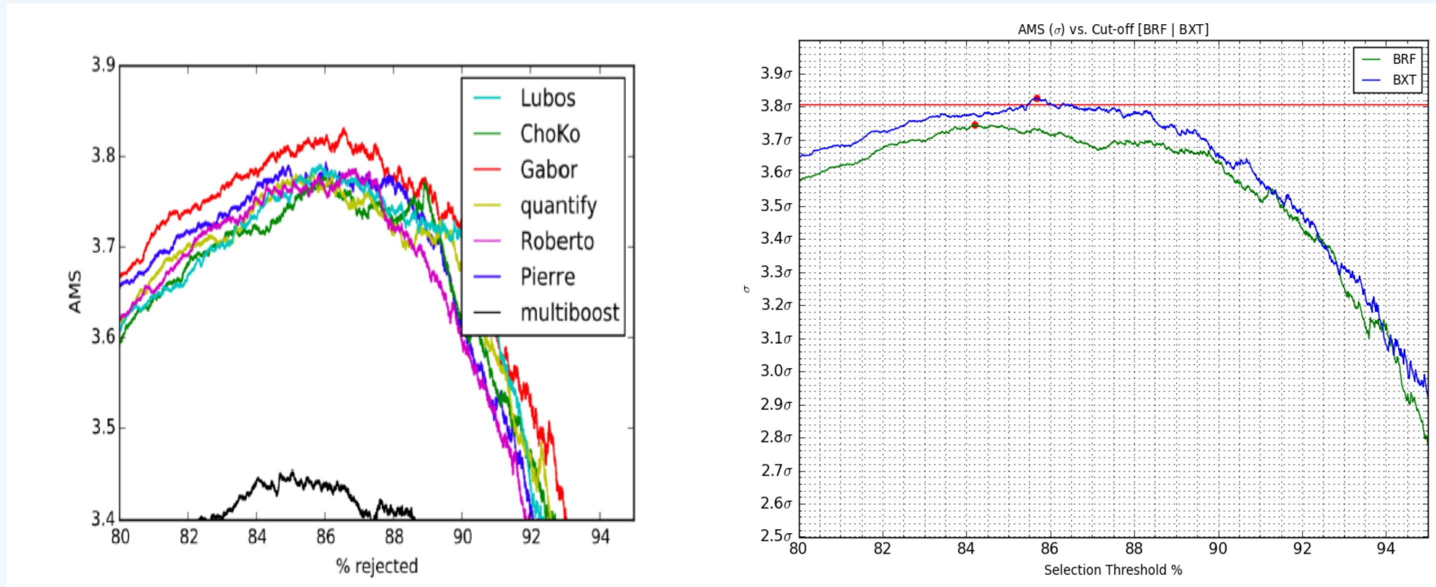


Figure 6: Leading solutions for the ATLAS Higgs classification (left) and the proposed algorithm (right) which uses a boosted version of tree ensembles. The green curve shows the significance curve obtained by boosting traditional forests (BRF) and the blue curve is obtained by boosting extremely random trees (BXT).

References

C. Adam-Bourdarios et al. "The Higgs Boson Machine learning challenge". *JMLR : Journal of Machine Learning Research.* Vol. 42. 2015.

Vidhi Lalchand

Department of Physics